INTERSTATE METRO TRAFFIC FINAL PROJECT

Ben Ahnen

Table of Contents

I. Executive Summary

In this report, I will identify predictors and other factors that influence the amount of West Bound interstate traffic volume between Minneapolis and St. Paul. I found this dataset from the UCI Machine repository. The dataset contains 48,204 rows of traffic information spanning six years, from 2012 to 2018. This data tracks whether or not it is a holiday, the temperature, rain per hour in millimeters, snow per hour in millimeters, percentage of cloud coverage, a short text description of the weather, a long text description of the weather, date and time collected in local CST, and reported traffic volume. To make this dataset easier to analyze, I manipulated the data to create multiple new columns. I converted the temperature from Kelvin to Fahrenheit so that a non-technical individual would be able to understand the information. I also normalized the temperature variable, rain per hour, snow per hour, and percentage of cloud coverage. I inserted a categorical holiday indicator (0 or 1), a categorical traffic volume level indicator (0 or 1), a day of the week column, day of the month column (1-31), a year column, and month column (1-12). My initial thought of building a regression model using the features on predicting the traffic volume led to the expected outcome that we would not be able to precisely determine the exact volume of traffic. Because of this, I pivoted my research to see if the traffic level (higher or lower than the median traffic volume amount) could be predicted.

There were multiple questions I wanted to answer while analyzing this data. In my exploratory data analysis phase. I wanted to see if the day of the week impacted the average amount of traffic. I also wanted to see if holidays tend to have heavier traffic or not. I was curious to see if average traffic volume varied by month. Lastly, I wanted to further investigate to see if the weather type impacted the average amount of traffic recorded. Once I had completed the exploratory analysis of the data, I created a generalized linear model (GLM) to see if certain facets of the data can be used to predict whether the traffic volume would be high or low.

I built tables and histograms to explore and gain a better understanding of the data. All the descriptive statistic tables below will include total traffic volume, mean traffic volume, median traffic volume, and the standard deviation of traffic volume. I will create four histograms with all of the graphs having the average traffic volume as the independent variable. The dependent will vary by histogram, however, and will be as follows: day of the week, holiday (IE: Christmas, Labor Day, none), month, and type of weather. For the GLM, I will calculate the precision, accuracy, recall, F1-score, and AUC of the model to test its effectiveness. Precision tests of the predictions made how many are accurate out of those predicted results. Accuracy tests the proportion of correct predictions out of all the predictions made. Recall tests of the predictions made how many are accurate of the actual class. F1-score is the average between the precision and recall. Lastly, the AUC checks how well the model is able to distinguish between

events and non-events in binary variables, such as if it is high or low traffic volume. Ideally, for all these tests, we want the value to be as close to 1 as possible.

I hypothesize that the weekend will have less traffic than the working week. I believe this because many people commute to work in the city during the week and my guess is that not all of them will go into the city on the weekend. I hypothesize that holidays will have a less heavier traffic flow than days without holidays. I assume that less people will be traveling on days they do not work and will be spending time with friends and families. For the next question I want to answer, I hypothesize that the average traffic by month will vary since Minneapolis gets cold. I think that the winter months will have a higher average because as it gets colder less people will want to be exposed to the elements. Lastly, I hypothesize that weather types will have some impact on the average traffic, however I believe clear skies will have the highest average because more people will want to be outside and traveling.

II. Methodology

To begin the process of exploring the data, I first looked for any extreme values to remove. Upon finding none, I moved into data manipulation and created new columns in the data frame. The original data set had nine columns of data. After the data manipulation was implemented, the data frame now had 20 columns. The columns that were added goes as follows: traffic_level, which adds a categorical indicator that shows whether the traffic level is above or below the median. Holiday_level, which adds a categorical indicator that shows if there is a holiday. Breaking up the data time column into month, day, hour, and year. Temp_fahrenheit, which transforms the data from Kelvin to Fahrenheit. Lastly, I added normalized columns that normalized the rain per hour, temperature in Fahrenheit, snow per hour, and the cloud coverage percentage. Adding these columns allowed me to analyze the data and make sure that the differing values were accounted for and able to be compared.

Once I had the data frame how I liked it, I began the exploratory analysis phase of the project. I started by grouping the data by day and outputting the total, median, standard deviation, and mean of the traffic by day. As seen below in *figure 1*, the mean traffic value stays relative steady during the week but drops quite a bit on the weekend. To better illustrate this, I have created a histogram that can be seen below in *figure 2*.

Figure 1

Descriptive Statistics For Days

Day	Total Traffic	Median Traffic	SD Traffic	Mean Traffic
Friday	24994869	4331.0	2025.19	3656
Monday	23403986	3623.5	2016.97	3309
Saturday	18946722	3003.0	1582.45	2774
Sunday	16276939	2260.5	1482.67	2369
Thursday	24799562	4280.0	2096.86	3638
Tuesday	23882653	4070.0	2099.00	3489
Wednesday	24831553	4315.0	2112.81	3583

Figure 2

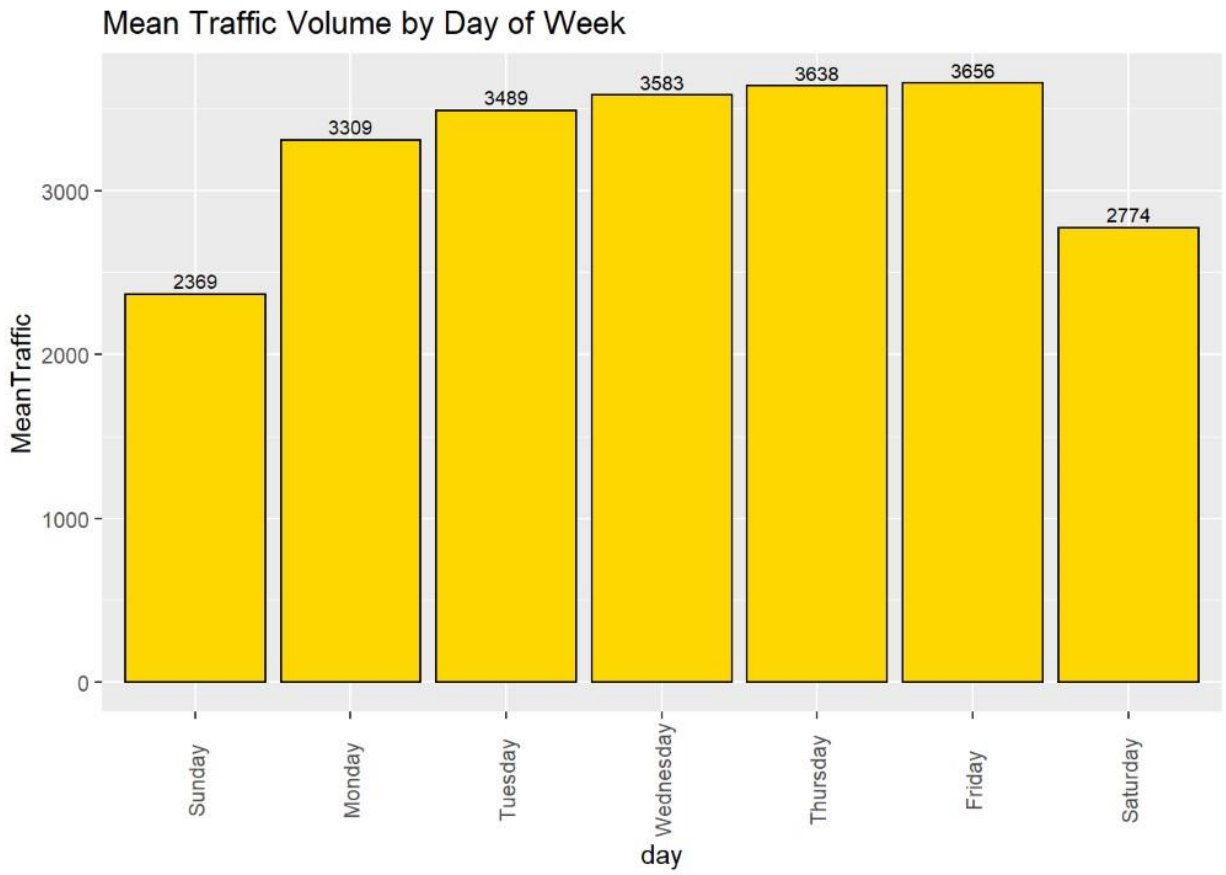


Figure 3 is the descriptive statistics table that captures the same statistics listed in *figure 1*, however it is grouped by holiday instead, and *figure 4* is the histogram of that table. The histogram shows the “none” holiday has a significantly higher average traffic volume as compared to days with holidays. As seen in *figure 4*, any holiday causes a drastic drop in traffic volume. I believe this is because anyone who travels on the holidays will be traveling before and after the holiday, not on the day of.

Figure 3

Descriptive Statistics For Different Holidays

Holiday	Total Traffic	Median Traffic	SD Traffic	Mean Traffic
Christmas Day	4965	767.5	145.99	828
Columbus Day	2597	494.0	63.37	519
Independence Day	5380	1060.0	100.86	1076
Labor Day	7092	1026.0	46.47	1013
Martin Luther King Jr Day	3676	600.0	80.80	613
Memorial Day	5538	1082.0	257.82	1108
New Years Day	8136	1458.5	277.11	1356
None	157083492	3385.0	1986.26	3263
State Fair	3174	655.0	31.55	635
Thanksgiving Day	5601	876.0	200.99	934
Veterans Day	3457	572.0	207.99	691
Washingtons Birthday	3176	623.0	88.92	635

Figure 4

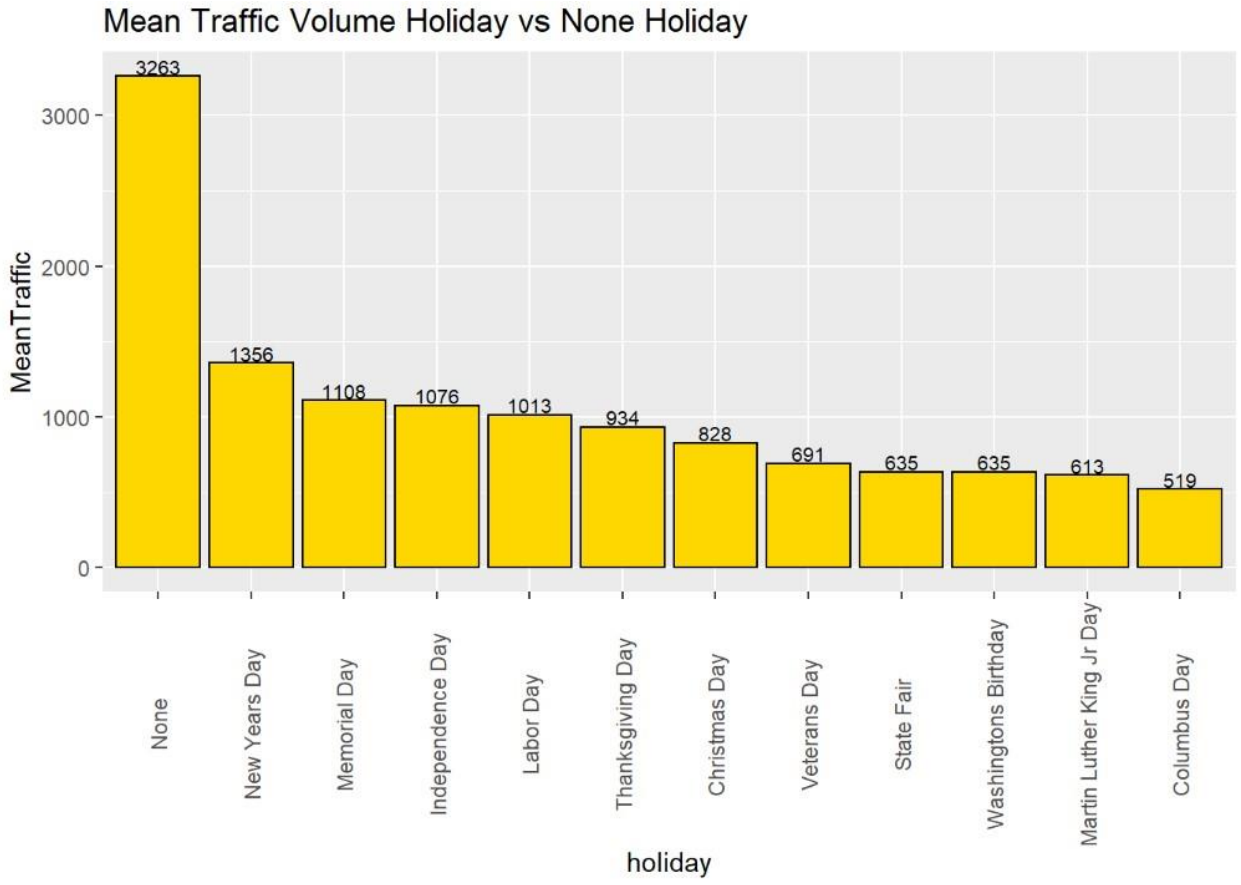


Figure 5 is the descriptive statistics table that captures the same statistics listed in *figure 1*, however it is grouped by months instead, and *figure 6* is the histogram of that table. The histogram shows months don't particularly impact the mean traffic amount,

they all stay between ~3000 and ~3400. As seen in *figure 6*, any holiday causes a drastic drop in traffic volume. I thought that the winter months would yield higher traffic levels as less people would want to drive when it got colder out, however that assumption was incorrect.

Figure 5

Descriptive Statistics For Different Months

Month	Total Traffic	Median Traffic	SD Traffic	Mean Traffic
April	14073322	3340.0	2069.51	3304
August	14859991	3648.5	1988.36	3394
December	12850072	3055.0	1877.19	3024
February	11275956	3306.0	1975.13	3198
January	12222632	3089.5	1908.01	3051
July	15370285	3268.0	1940.09	3205
June	12896760	3652.5	1974.98	3419
March	12548718	3510.0	2027.31	3308
May	14932993	3463.0	2032.46	3366
November	11675747	3149.0	2003.65	3168
October	11775826	3713.0	2003.30	3391
September	12653982	3477.0	1997.36	3303

Figure 6

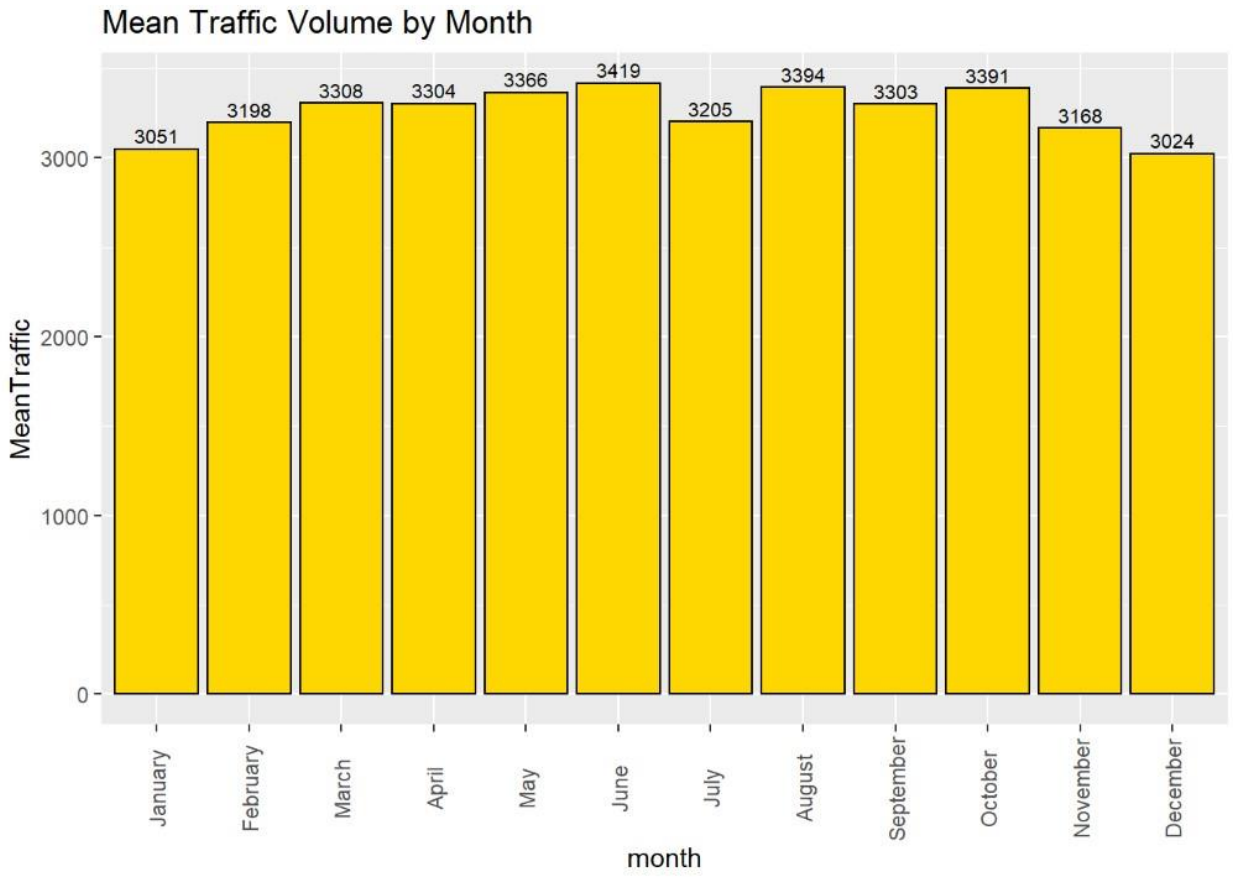


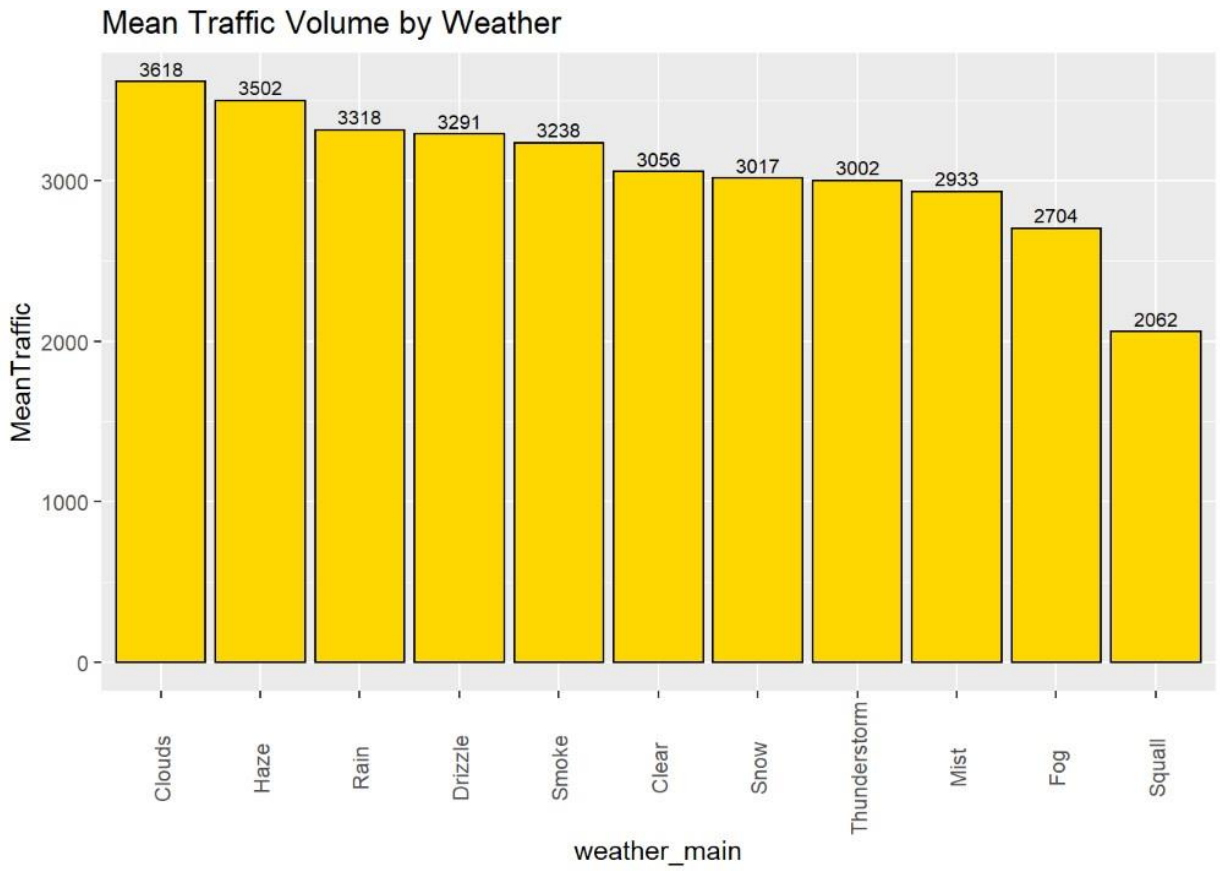
Figure 7 is the descriptive statistics table captures the same statistics listed in *figure 1*, however it is grouped by weather type instead, and *figure 8* is the histogram of that table. The histogram shows that the weather type impacts the mean traffic amount. I hypothesized that clear weather conditions would yield the highest mean traffic level, but that hypothesis was incorrect. I believe this to be due to the fact that people want to be outside when it is nice and will instead opt for biking, walking, etc. as opposed to traveling to an indoor location.

Figure 7

Descriptive Statistics For Different Weather Types

Weather Type	Total Traffic	Median Traffic	SD Traffic	Mean Traffic
Clear	40921675	3045.0	1987.10	3056
Clouds	54870172	4072.0	1906.20	3618
Drizzle	5992414	3473.0	1997.90	3291
Fog	2465793	2339.0	2125.53	2704
Haze	4762858	3987.0	1873.51	3502
Mist	17451092	2800.0	2073.03	2933
Rain	18819160	3497.5	1982.23	3318
Smoke	64753	3612.0	1978.02	3238
Snow	8676444	3080.0	1900.19	3017
Squall	8247	1818.0	1950.07	2062
Thunderstorm	3103676	2971.5	1988.30	3002

Figure 8



In order to properly evaluate the GLM, I needed to find which variables I thought would be a key indicator of traffic volume. I settled on the following variables: holiday_categorical, norm_fahrenheit, hour, weather_main. I used the facets of the data to predict the categorical traffic level, whether it would be above or below the median

traffic level. I will discuss my findings on this model further in the Analysis & Discussion section below.

III. Analysis & Discussion

The Descriptive statistics and visualizations provided a basic understanding of the features and how they are significant with respect to the traffic volume. I proceeded in my analysis by developing a GLM using the holiday_categorical, norm_fahrenheit, hour, and weather_main columns. With this model I was hoping to identify if these features were able to predict if the traffic level was high or low.

The data is split into training and test data using random sampling with a split of 70:30 with the training data used to train the model and test data for evaluating model performance. Using the training and testing datasets created, I proceeded to develop a GLM. The predicted output ranged from values of zero to one. Because of this, I assigned predicted scores >= .5 as “1” indicating a high level of traffic, and scores of < .5 as “0” indicating a low level of traffic. I tested the effectiveness of the model by evaluating the model performance based on accuracy, precision, sensitivity, F1-score, and AUC. The scores were all tracked in a table and can be seen in *figure 9* below.

Figure 9

	Accuracy	Precision	Recall	F1	AUC
GLM	0.908	0.88	0.944	0.911	0.9081

The goal of the evaluators is to have them as close to a score of 1.0 as possible. Because of this, we can see that overall, the model was able to accurately predict a vast majority of the categoric traffic level. While these tests do a great job of showing how well the model performs, I wanted to see where the model was overestimating. *Figure 10* represents the total amount of instances in which the testing dataset had a high level of traffic. *Figure 11* represents the total amount of instance in which the testing dataset had predicted a high level of traffic.

By comparing *figure 10* and *figure 11* below, it is evident that the model tends to predict high traffic volume slightly more than high traffic volume occurs. Every type of weather was overstated by the model in terms of the total amount of times that the particular weather type was predicted to have a high level of traffic. On a percentage basis, the largest discrepancy was smoke, which was predicted to have a high level of traffic three times more than it did. In terms of raw numbers, mist had the largest difference at 151 more predicted instances than what actually occurred.

Figure 10

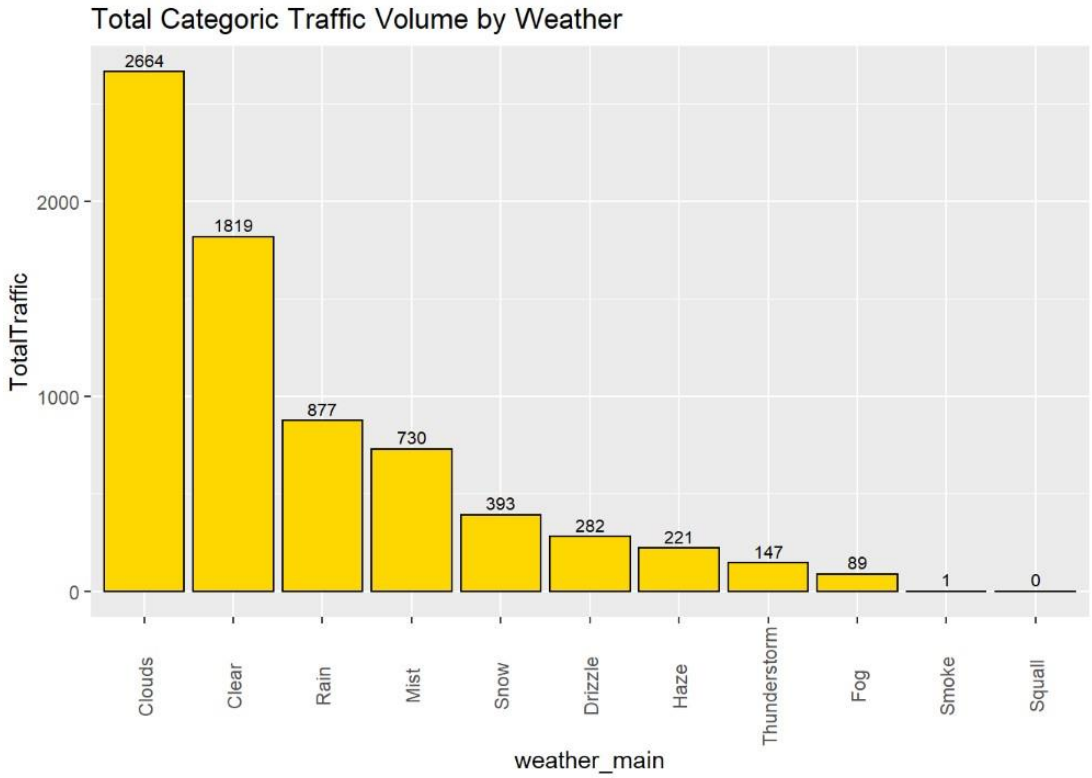
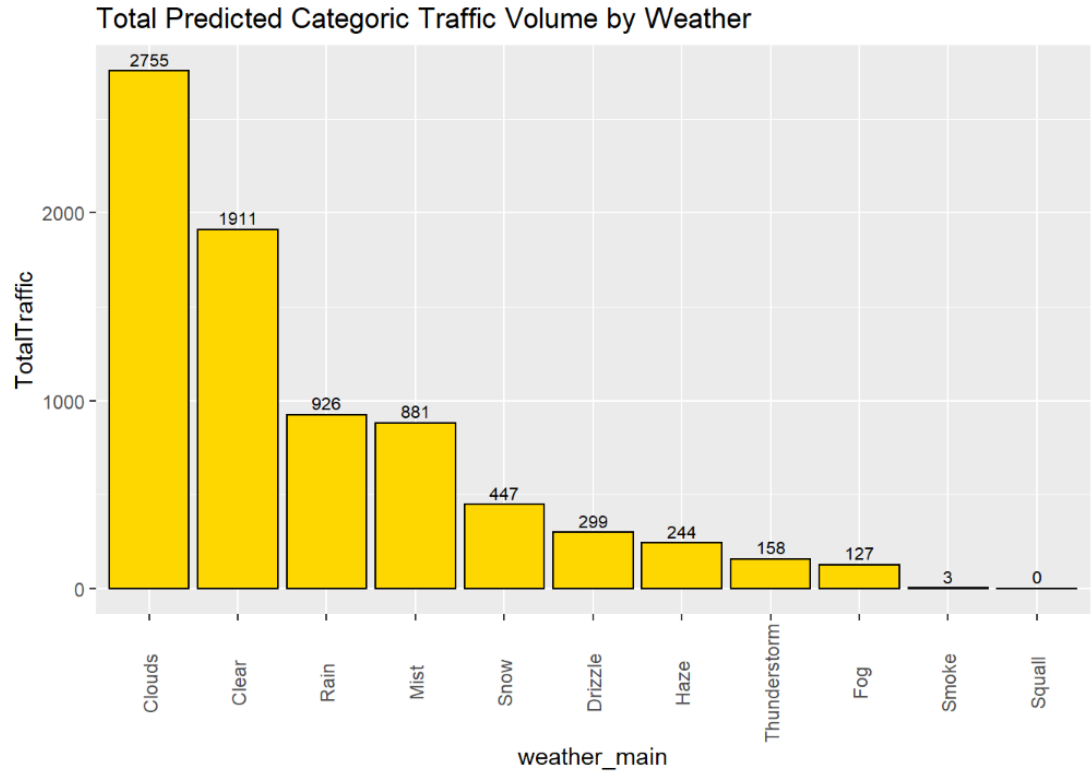

Total Categoric Traffic Volume by Weather
TotalTraffic
2000
1000
0
2664
1819
877
730
393
282
221
147
89
1
0
Clouds
Clear
Rain
Mist
Snow
Drizzle
Haze
Thunderstorm
Fog
Smoke
Squall
weather_main

Figure 11


Total Predicted Categoric Traffic Volume by Weather
TotalTraffic
2000
1000
0
2755
1911
926
881
447
299
244
158
127
3
0
Clouds
Clear
Rain
Mist
Snow
Drizzle
Haze
Thunderstorm
Fog
Smoke
Squall
weather_main

Lastly, I looked at the model assumptions. *Figure 12* shows the model assumptions for the Residuals vs. Fitted. The Residuals vs. Fitted should be more or less evenly distributed across the data, and they should also have a mean of 0. The model is on the fringe of meeting these assumptions. The data has some gaps but is more or less evenly distributed among the Y-axis and encompasses about two thirds of the x-axis. As for the mean remaining at 0, the red line (mean) remains very close to 0. For these reasons, I believe that the model meets the model assumptions for Residuals vs Fitted but could be better. *Figure 13* shows the model assumptions for the Normal Q-Q. The Normal Q-Q should follow the line closely. However, as *figure 13* shows, the data does not follow the black dotted line closely. For this reason, I would say that it does not meet the model assumptions for the Normal Q-Q plot.

Figure 12

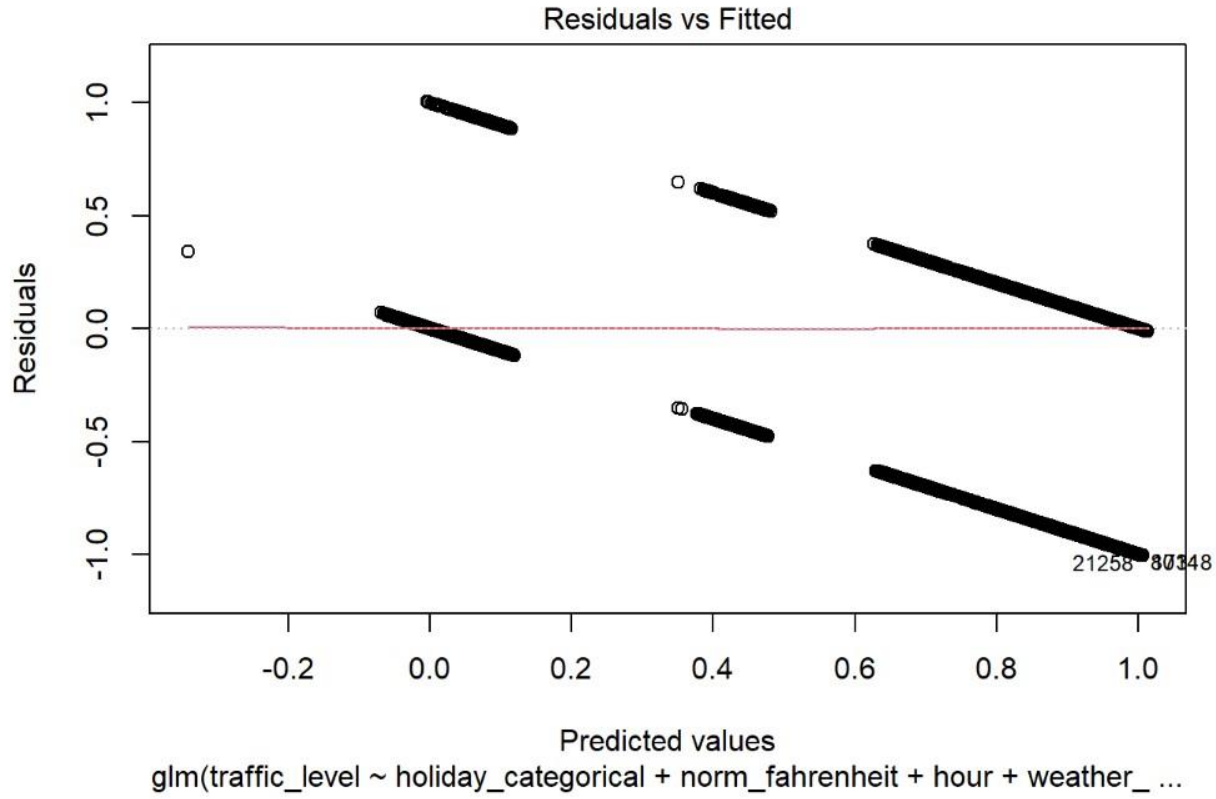
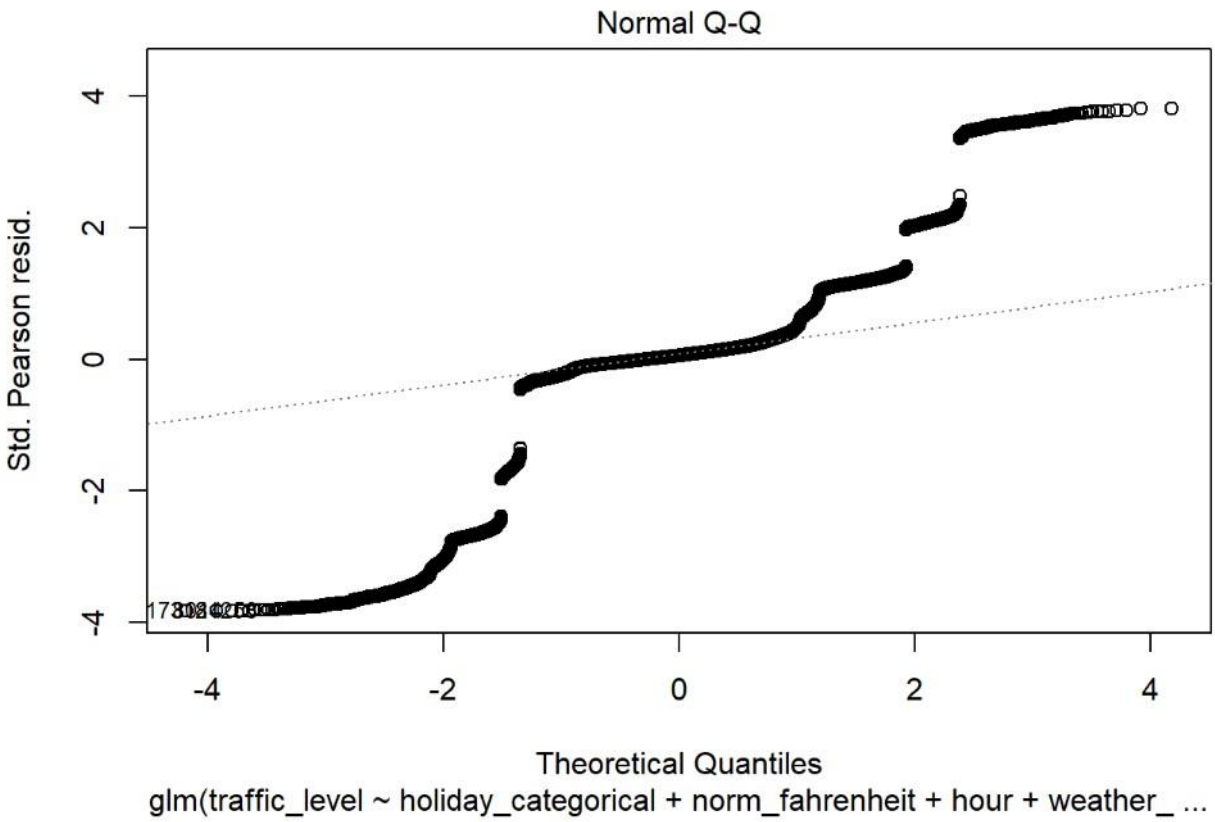


Figure 13


Normal Q-Q
Std. Pearson resid.
4
2
0
-2
-4
-4
-2
0
2
4
Theoretical Quantiles
glm(traffic_level ~ holiday_categorical + norm_fahrenheit + hour + weather_ ...

IV. Conclusion

Based on my research, I have concluded that hour of day, weather type, whether or not there is a holiday, and the temperature outside are effective predictors when trying to model traffic level. The model tends to predict that traffic levels will be higher than they are, which could impact the implementation of this model. This model could be used to help build construction schedules or when a company wants to advertise on a billboard along the interstate. With the model running at about 91% accuracy, I would say that it would not have major implications but should be considered upon implementation. Based off the model assumptions, I would say that further modeling needs to be done. It can be fine tuned a bit so that it meets more of the model assumptions and produces more accurate results.

Some general limitations that I encountered are that there are no traffic construction columns available. If this data was included in the dataset, I would eliminate instances when there was construction near or on that road. I would do this so that I could run on the assumption that I could accurately predict busier times and plan an accurate construction schedule for the future. Additionally, another limitation of the data is that it is from 2012-2018. It does not reflect the current climate of many people working from home. Due to this, I assume that traffic levels during the week would be less than is shown in the model as less people are traveling into the office.

V. References

UCI Machine Learning Repository: Metro Interstate Traffic Volume Data Set. (n.d.). Retrieved May 1, 2022, from https://arcHive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume